Naval Research Laboratory

Washington, DC 20375-5320

NRL/FR/5531--93-9569

Voice Message Systems for Tactical Applications (Canned Speech Approach)

G. S. KANG
T. M. MORAN
D. A. HEIDE

Human Computer Interaction Laboratory Branch
Information Technology Division

September 3, 1993

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave Blank*)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 3, 1993	Continuing — 01 Oct. 91 - 30 Sept. 92

4. TITLE AND SUBTITLE

Voice Message Systems for Tactical Applications
(Canned Speech Approach)

5. FUNDING NUMBERS

PR - 33904N
61153N

6. AUTHOR(S)

George S. Kang, Thomas M. Moran, and David A. Heide

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
Washington, DC 20375-5320

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/FR/5531--93-9569

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Space and Naval Warfare Systems Command
2451 Crystal Drive
Alexandria, VA 22245-5200

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

One-way noninteractive voice messages are often used in tactical environments. Examples are surveillance and reconnaissance reports, tactical coordination messages, warnings, and reminders. These messages can be transmitted efficiently in terms of words, phrases, or sentences. We use a speech recognizer to convert speech into text. The resultant data rate is below 100 b/s.

At the receiver, speech is regenerated by concatenating the stored speech waveforms (canned speech) corresponding to the received indices. Intelligibility of the resultant speech is high because output speech stems from actual stored speech rather than synthetic speech. If tactical messages are generated by concatenating only words, however, we need to incorporate the sentence-level prosody in the generated speech. For tactical messages, however, prosodic rules are relatively simple because tactical messages are customarily spoken without significant pitch and rhythmic inflections. If tactical messages are generated by concatenating phrases and sentences, resultant speech will sound natural without further speech modification. Another advantage of the "canned speech" approach is that the spoken language can be translated into any one of the preselected languages at the receiver. The voice message system will play a significant role in future DoD secure voice terminals.

14. SUBJECT TERMS

Tactical voice message
Formatted voice messages
Extreme-low-data-rate voice messages

15. NUMBER OF PAGES

30

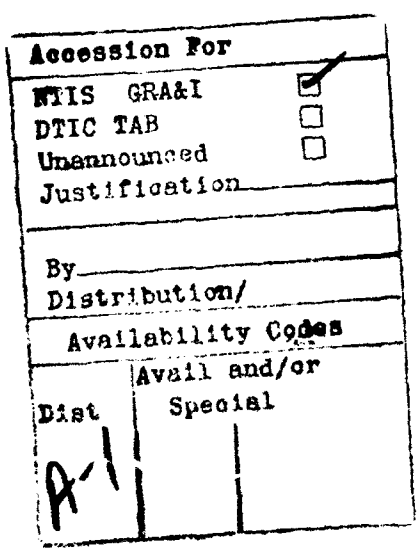
16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std 239-18
298-102

CONTENTS



VOICE MESSAGE SYSTEMS FOR TACTICAL APPLICATIONS (CANNED SPEECH APPROACH)

INTRODUCTION

After the August 1990 Iraqi invasion of Kuwait, ship Naval messages reached an average daily rate of 1500. Says Radioman Chief Charles Sawtelle, "The system was so clogged that it couldn't get mail to the crew in much less than 10 days." To alleviate the problem, messages were labeled: O for immediate, P for priority, and R for routine. Messages labeled R were withheld until they could be delivered without hindering the delivery of the others [1].

The Navy is trying to change this inefficiency. Through the Copernicus program of VADM Jerry Tuttle, the director of the Navy's Space and Electronic Warfare division at the Pentagon, the Navy envisions efficient information transfer systems that rely on commercial off-the-shelf, open-architecture multimedia systems [2]. Under Copernicus, "preemption," "prioritization," or "selective withholding" will not be the solution to declog information transfer systems. The solution will be to use high-volume datalinks on one hand and data-rate compression on the other hand. Accordingly, we at the Voice Systems Section of the Naval Research Laboratory are performing three R&D tasks related to voice data rate compression:

1. *800-b/s and 1200-b/s Voice Encoders:* These voice encoders are for two-way interactive voice communication similar to the telephone. Functionally, these voice encoders are similar to the currently widely deployed Secure Terminal Unit (STU-III) or Advanced Narrowband Digital Voice Terminal (ANDVT), both operating at 2400 b/s. These R&D efforts were documented in two NRL reports [3,4]. Currently, we are using these voice encoders in the implementation of the Narrowband Multimedia Terminal capable of transmitting voice and data (images, overlays, simple hand-drawn sketches, indices of stored maps, etc.) simultaneously over narrowband links.

2. *Voice Message System:* A voice message system is for one-way information transfer like e-mail. Since immediate interaction is not involved, voice messages can be sent in nonreal-time. One type of voice message is free-form where the vocabulary is limited but text is unconstrained. We transmit such a message in terms of the indices of the individual words. At the receiver, speech is generated by concatenating the raw speech waveform of individual words stored in memory (Fig. 1). The data rate is extremely low (i.e., below 100 b/s, even for a vocabulary size of 10,000 words). The word intelligibility is excellent because no analysis/synthesis (i.e., vocoding) is involved. Words, however, must be strung together properly to generate natural sounding sentences. This is one of the topics discussed in this report.

3. *Formatted Voice Message System:* We also investigated the transmission of voice messages from a fixed format. The data rate required to transmit formatted message is far less than that

Manuscript approved May 19, 1993.

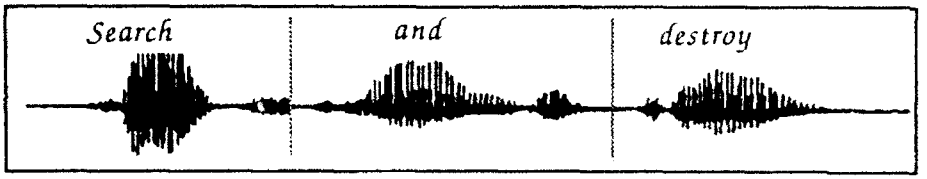


Fig. 1 — Compressed speech waveform of a brief tactical message generated by concatenating raw speech from individual words. In the past, such an approach has been implemented and used for limited applications, but design procedures have not been documented. In this report, we present the technique in some detail.

required to transmit free-form messages (i.e., far below 100 b/s) because speech is transmitted in terms of words, phrases, and sentences. Again, we concatenate canned speech waveforms to generate continuous speech. The resultant speech sounds natural because words, phrases, and sentences are naturally spoken. We have implemented the Formatted Voice Message System operating in the LINK-11 environment [5]. By using the Formatted Message system, the voice coordination message can be transmitted over the data link, rather than a separate and dedicated voice link as currently required. The Formatted Voice Message System is also a topic of this report.

For many years, our R&D efforts have been focused at improving speech intelligibility of two-way interactive voice communication at various data rates, from 800 to 16,000 b/s; Refs. 3 and 4 are examples. In this report we are concerned with improving speech intelligibility of one-way noninteractive tactical voice messages at data rates below 100 b/s.

This report was written for three groups of people: program managers and sponsors who are actively involved in the transfer of voice technology to working hardware; communication-architecture planners who are interested in the state of the art of voice transmission; and independent researchers who develop voice terminals. We hope that this report provides some useful information to these individuals.

BACKGROUND

The majority of tactical voice communication is transmitted at a low data rate to attain the following desirable features: antijam, low probability of detection, longer transmission distance for a given transmitter power (and antenna gain). We can use much lower data rates to transmit one-way voice messages because of the characteristics mentioned below.

Tactical Voice Messages

We may divide tactical voice communication into two different types: two-way interactive voice communication, and one-way noninteractive voice communication. There are significant differences between the two.

Two-Way Voice Communication: Since the invention of the telephone, two-way interactive dialogue is a widely used mode of communication. Two-way voice communication achieves speedy information forward as well as a speedy response backward. Speaker recognizability from spoken voice

is important because people tend to be reluctant to talk unless the speaker knows who is at the receiver. We do not attempt to generate interactive dialogue by canned speech because interactive dialogue requires a large vocabulary, complex speech patterns, and swift exchanges. We use a low-data-rate (800 or 1200 b/s) voice encoder [3,4] for two-way interactive voice communication.

One-Way Noninteractive Voice Communication: Another mode of voice communication is non-interactive one-way voice message transfer. One-way messages usually do not require an immediate verbal response from the listener. Characteristics of one-way voice communication are significantly different from those of two-way communication:

- *Brevity:* Messages are generally brief. They include surveillance reports, tactical coordination information, warnings, and reminders. Examples are: "No enemy sighted," "Maintain radio silence!," "Cease fire!" etc.

- *Jargonization:* Tactical messages are often jargonized to make messages more easily understandable in poor signal-to-noise conditions. Examples are voice coordination messages over Navy LINK-11. In this case, messages are even shorter.

- *Limited Vocabulary:* Vocabulary is limited because messages are related to the execution of specific tasks.

- *Speaker recognizability:* Speaker recognition from spoken voice is not essential. Frequently, tactical messages are broadcast over the air, and listeners do not care about who is actually talking.

- *High Articulation:* Tactical communicators tend to articulate each word to make messages more easily understood. Words are often spoken in isolation without pitch inflections.

This report is concerned with transmission of one-way voice messages. Our objective is to transmit one-way voice messages at extremely low data rates (below 100 b/s) by not using the conventional two-way voice communication techniques. The difference between two-way voice communication approach and our message transmission approach will be compared in the following paragraphs.

Sound Transmission vs Message Transmission Devices

Two-way voice communication devices (such as the telephone, walkie-talkie, CB radio, etc.) are *sound transmission* devices. They are designed to reproduce the speaker's accent, rhythm, pitch contours, and vocal timbre. The same can be said for the current digital telephones, particularly *waveform encoders* used in Government secure voice systems (Fig. 2). Low-data-rate voice encoders such as the 2400-b/s linear predictive coder (LPC) are also capable of reproducing (although less faithfully) the speaker's accent, rhythm, pitch contours, and vocal timbre.

In one-way tactical message transmission, however, the conveyance of information is the critical part. If only the message is transmitted in terms of words, phrases, or sentences, the required data rate is much lower, below 100 b/s . The encoded bit stream, however, must be heavily protected against transmission bit errors, because a single error in the encoded data could affect regenerated speech much more adversely than vocoded speech.

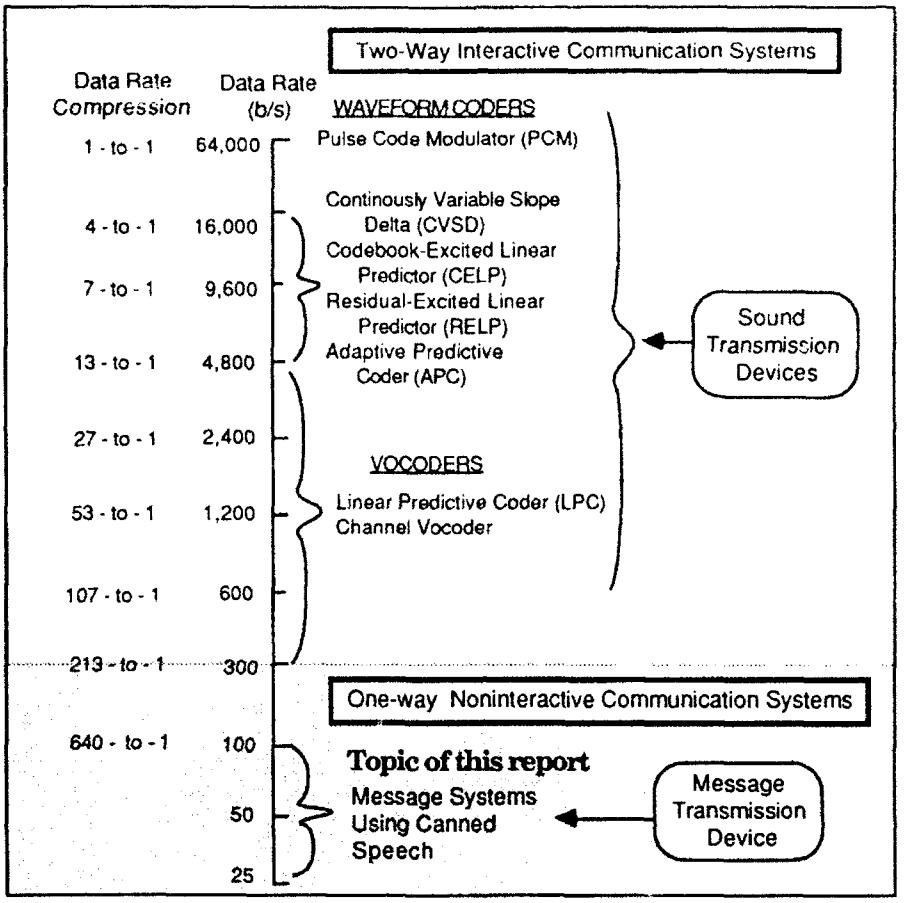


Fig. 2 — Various means of voice communication. In the sound transmission device, much of the data rate is consumed by the transmission of the speaker's accent, rhythm, pitch contours, and vocal timbre. These features are absent in the message transmission device that permits an extremely low transmission rate.

It is significant to recognize that when a sound transmission device is used, acoustic noise interference affects speech intelligibility and quality. For example, the Diagnostic Rhyme Test (DRT) score of the 2400-b/s LPC with F15 background noise is only 69 (which is unacceptably low). On the other hand, if a message transmission device is used to transmit voice message, and messages are entered into the transmitter by a nonverbal means (i.e., keyboard or mouse), acoustic interference does not affect the output speech quality. Immunity to noise interference is an advantage of the voice message system.

Why Voice Output?

Since a string of words is transmitted, received messages can be read. People often ask why do we need voice output? Here are some of the reasons:

- *Urgent Messages:* Certain messages require immediate broad dissemination. Examples are warning messages (e.g., inbound missile, fire, explosions, etc.). These messages are best heard from the intercom because everyone must act immediately.

- *Unattended Terminals:* A response to a text message must await the communicator reading the text. A delay of message reading could be a disaster if the message is a warning of impending disaster. Voice output over an intercom can instantly notify the appropriate recipients.

- *Emotional Content*: Speech output can be more effective than written output, because speech can exhibit the emotional content of the message. When a message is spoken loud and fast, we tend to think that the message is more urgent.

- *Speech is Preemptive:* When we hear speech, we tend to shut off other perceptive stimuli (e.g., vision). Thus, speech is a more effective means for delivering tactical messages, because we tend to pay more attention. An exception would be for those working in an environment where speech sounds are heard continuously (as encountered in the Combat Information Center).

- *Eyes are Busy:* Often, communicators perform multiple duties. A high-performance aircraft pilot is a communicator, navigator, and missile deployer. When a communicator performs multiple tasks, speech is a preferred means of communication, since it permits the use of the eyes for other visual needs.

- *Faster Message Transfer*: We can understand speech spoken at a faster rate than what is normally heard [6]. Thus, speech can be played at a faster rate to reduce the information transfer time.

- *Annotation of Visual Display:* Images (e.g., maps, photos) and related annotations are increasingly being transmitted together over networks. Annotations of visual displays can be best accomplished by speech rather than text. In this way, eyes can be focused on the image, while the person is listening to the annotation.

TECHNICAL APPROACH

Figure 3 is a block diagram of the voice message systems. Through a visually-prompted menu, the computer guides the user in selecting an appropriate choice of information to be entered in the input interface. There are three significant factors of the extreme-low-data-rate tactical voice message system (see the thick-lined blocks in Fig. 3). They are: (1) message transmission in terms of words and phrases, (2) speech regeneration by concatenating canned speech (i.e., no synthetic speech), and (3) speech recognizer as input interface. We will discuss each aspect.

Message Transmission in Terms of Words and Phrases

If speech is directly quantized, the data rate is on the order of 64,000 b/s. If the speech waveform is decomposed to its spectral envelope and pitch harmonics (Fig. 4), the speech data rate may be reduced to 4,800 to 16,000 b/s. Simplification of pitch harmonics further reduces speech data rate to 2,400 b/s. In these approaches. speech parameters are transmitted as often as 50 times per second. On the other hand, in the voice message system, speech is transmitted in terms of words or phrases. Since we speak approximately two words a second or one phrase every few seconds, the resultant data rate becomes extremely low, somewhere below 100 b/s.

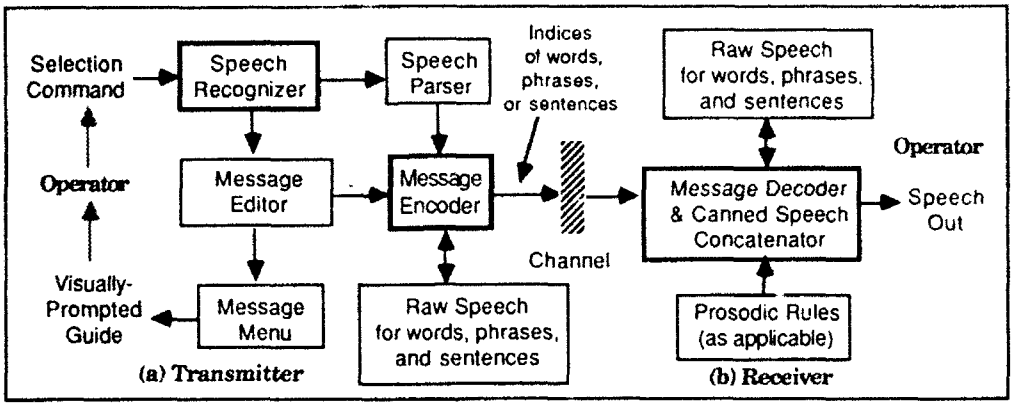


Fig. 3 — Simplified block diagram of Voice Message System. The thick-lined blocks represent new technical approaches for the Voice Message System presented in this report.

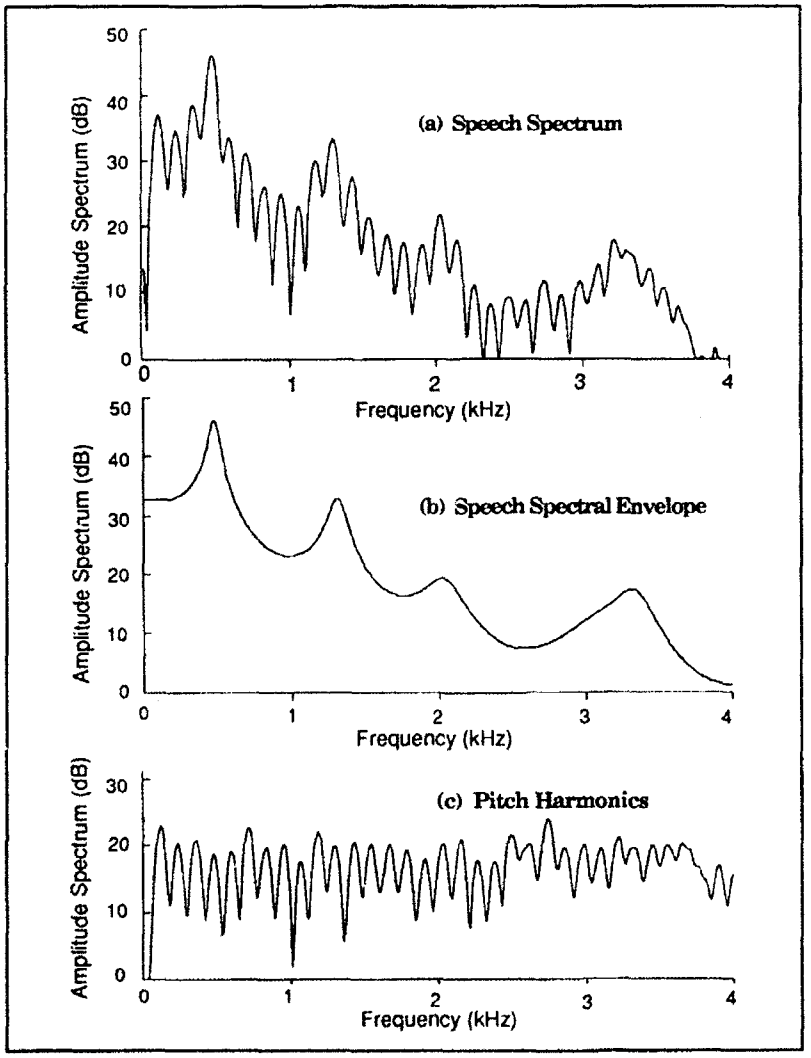


Fig. 4 — Speech spectral decomposition. In speech encoding, the speech spectrum (Fig. 4a) is often represented by a product of the speech spectral envelope (Fig. 4b) and pitch harmonics (Fig. 4c).

Speech Regeneration by Concatenating Speech Waveforms

When voice messages are transmitted in terms of words, there are two possible ways of converting a string of words into speech. One is by using a text-to-speech converter, and the other is by using canned speech. We prefer the canned speech approach because of the reasons discussed below.

Synthetic Speech Approach

A number of text-to-speech converters are currently available from commercial sources. DECtalk, a product of Digital Equipment Corporation, is a notable example [7]. We are certain that a text-to-speech converter could be used for certain applications, but we do not think it is good enough for the generation of tactical voice messages. We list five significant reasons why a text-to-speech converter cannot be used for our purpose:

- *Low Intelligibility:* Speech generated from text is synthetic speech generated from a limited number of spectral parameters and an artificial excitation signal. Such a simple speech model generates degraded speech. Repeated hearing of synthetic speech often does not improve comprehension because the speech spectrum lacks fundamental acoustic cues. Even a book published by the manufacturer of DECtalk states that "The increased difficulty of listening to synthetic speech makes it of dubious value in warning and emergency systems." [8].

- *Susceptibility to Noise Interference*: Resonant frequencies of synthetic speech are not as sharp as those of natural speech (compare Fig. 5a with Fig. 5b). As a result, synthetic speech quickly loses intelligibility at noisy sites.

- *Poor sound quality*: Synthetic speech is generated by a filter driven by an artificial excitation signal (pulse train to generate vowels or random noise to generate consonants). This simple speech model is incapable of producing high-quality speech. The speech quality evaluated by the Diagnostic Acceptability Measure (DAM) of synthetic speech is only in the low 50s, whereas live speech is in the 80s.

- *Incorrect pronunciation of certain words:* A text-to-speech converter uses a set of pronunciation rules to generate speech from text. Apparently, rules do not cover all cases. Certain words are pronounced incorrectly.

- *Unnaturalness:* Says Klatt, the inventor of DECtalk, "Why doesn't DECtalk sound more like my original voice, after years of my trying to make it do so? According to the spectral comparisons, I am getting pretty close. But there's something left that's elusive, that I haven't been able to capture. It has been possible to introduce these details and to resynthesize a very good quality of voice. But to say 'here are the rules, now I can do it for any sentence' — that is the step that failed miserably every time." [9].

Canned Speech Approach (Our Choice)

The best approach for generating the highest quality speech is to transmit raw speech, but this approach would be unacceptable by narrowband users because the required data rate is too high (64,000 b/s). Instead, we divide speech in terms of words, phrases, or sentences; and we store their speech waveforms in memory. We concatenate these speech waveforms. The advantages are:

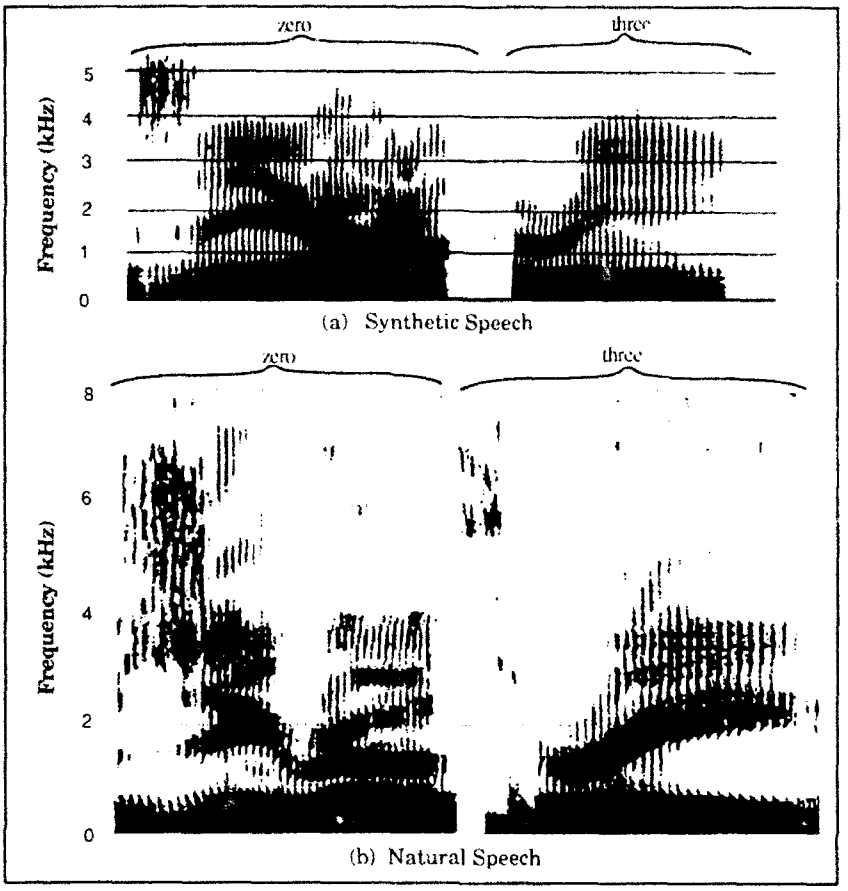


Fig. 5 — Spectrograms of synthetic speech (from a text-to-speech converter) and concatenated raw speech. Synthetic speech has a limited bandwidth, and its formant structure is not as clear as that of raw speech

- *High Intelligibility and Quality*: In the unformatted voice message system, word intelligibility is high because each word is represented by raw speech, not synthetic speech. Not only is word intelligibility high, but word quality is also high. The stress and rhythm within a word is proper because each word is represented by the speech waveform of an actual spoken voice. In the formatted voice message system, speech intelligibility and quality are even better than those of the unformatted voice message system because words, phrases, and in some cases complete sentences are naturally spoken voice.

- *Automatic Language Translation Capability:* Each word can be translated to another language by simply reading from a different speech waveform library. Among certain languages, a sequential translation of words leads to an idiomatic translation of the language. Among other languages, however, such is not the case. But, for brief tactical massages (e.g., "cease fire," "attack immediately," etc.) the word-by-word translation is sufficient to convey the meaning of the message correctly. Such a capability is beneficial for joint operation by a multi-national task force.

- *No possibility of mispronunciation*: By the same token, names of persons or places (which are often mispronounced by a text-to-speech converter) are pronounced correctly.

If speech is transmitted in terms of words (i.e., unformatted voice message system), we must incorporate some amount of prosodic details in the concatenated speech to make speech sound more natural. This is not easy to do because speech characteristics (utterance rate, pitch rate, shortening speech waveform, lengthening the speech waveform, etc. are difficult to accomplish). We will discuss this aspect in detail. On the other hand, if we transmit speech waveform in terms of phrases (e.g., formatted message system), the problem becomes simpler.

Speech Recognizer as Input Interface

The most natural means of inputting information is by natural voice using a voice recognizer. The user selects a message by saying the title into a speech recognition computer. After the computer shows that it has recognized the correct message, the user can tell the computer to transmit the message. The computer provides either sound or visual feedback to the user. Although the speech recognizer is a convenient form for inputting information into the computer, it must work with high accuracy to be accepted for tactical use. Recently we performed some tests and observations in a shipboard Combat Information Center (CIC) to evaluate the application of speech recognition technology. Here are some issues of speech recognition in general and shipboard environment in particular:

- *Ambient Noise:* Speech recognizers do not work well in a noisy environment. Noise cancellation would certainly be required. We found about 70 dB of constant background noise due to equipment (cooling fans, buzzers, etc.) and much louder intermittent noise due to whistles and voices on intercoms. We were told that the background sound levels in the CIC of this ship were lower than most.

- *Speaker Training:* Some recognizers require training to the user's voice. Generally, this is not a serious problem, as only a few people on board perform particular communication tasks.

- *User feedback:* This is an important concern. Operators in the CIC sit at consoles with many controls and large displays. They usually have their hands and eyes full. Any feedback method must have minimal impact on an operator's other duties. In the future, consoles are likely to be computer workstations with graphic windows. It may be possible to use a window as the interface to provide the feedback. The complexity of this interface varies with the quality of the speech recognizer, which, in turn, partially depends upon the number of messages. If a speech recognizer has a consistently high error rate, the remainder of the interface must provide the user with the option of correcting recognition mistakes or provide alternative input methods.

- *Microphone Placement*: Speech recognizers tend to be extremely sensitive to microphone placement. The user must carefully maintain the microphone in the optimum position.

- *Vocabulary Size:* Many speech recognizers claim to have a large vocabulary (10,000 words or more). While others have only a thousand words. The grammar serves to structure the stored word patterns with which the recognition is performed. This structure limits the vocabulary at any one time to a small portion of the whole. Some recognizers model the grammar according to the natural language, i.e., English. It is possible to create special grammars for particular situations. In any event, some structuring of the vocabulary is necessary, so vocabulary size is not an important issue as long as the recognizer works well.

- *Continuous Speech vs Isolated Words:* Some recognizers are able to parse words from continuous speech. Others can only recognize a discrete utterance (word or phrase). The former is easier and more natural to use. The latter generally has higher recognition accuracy. For this reason, if the vocabulary is quite small (less than 50 words), discrete utterance recognition might be preferable, especially in a noisy environment.

It is convenient to separate voice message systems into two categories. One is the *free-form (or unformatted) message system* where the vocabulary is limited, but text is unconstrained. The other is the *formatted voice message system*. Formatted messages may be divided into three categories: (1) the entire message is fixed (warning messages, reminders, etc.) (2) the message form has blank space where information is entered (reconnaissance and status reports), and (3) the messages are made of a collection of fixed words, phrases, and sentences. We will discuss these two types of voice message systems.

UNFORMATTED VOICE MESSAGE SYSTEM

The allowable vocabulary is limited, but there is no restriction on text. At the transmitter, the individual word is entered into the computer by one of three possible means: natural speech via a speech recognizer, printed text via an optical character reader, and keyboard input (Fig. 6). The selected words are then converted to their indices. At the receiver, the speech waveform of each word is stored in the Speech Waveform Library. The words were preselected for each application, and they were read by a person with a good voice. The speech waveforms are carefully edited prior to storage. The compiler introduces appropriate prosodic details.

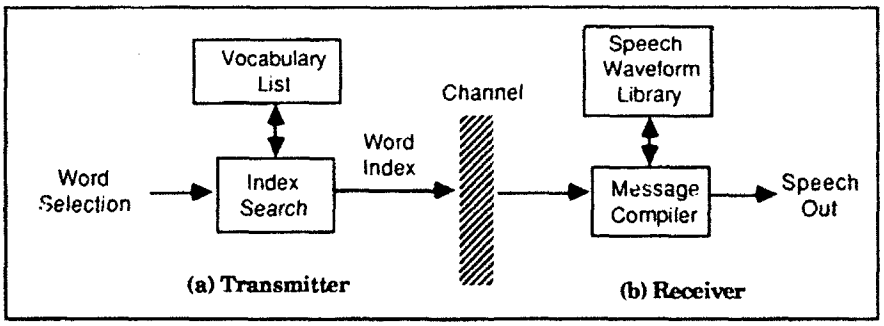


Fig. 6 - Essential elements of an Unformatted Voice Message System. This block diagram does not include the input interface that was discussed previously.

Certainly, such an approach is not suited for generating casual conversations where a slight shift in rhythm, intonation, or stress could alter the subtle meaning of the speech. But tactical voice messages are not casual speech. Messages are generally short and simple. Furthermore, military communicators are trained to articulate each word in isolation without much pitch inflection, so that messages are more easily understandable even in poor signal-to-noise conditions. In extremely noisy platforms, the communicators simply yell each word. Therefore, actual spoken tactical messages tend to sound like canned speech. Recently, we performed an experiment to generate tactical messages by the canned speech approach. This report discusses the lessons we learned through this experiment.

Speech Waveform Library

The Speech Waveform Library contains speech samples of individual words. The following factors are critical in the design of the speech waveform library:

Speech Bandwidth Selection

It is a well-known fact that speech intelligibility is affected by the speech bandwidth [10]. The conventional front-end bandwidth is 0 to 4 kHz, which is adequate for the reproduction of vowels. But it is too limited for voiceless fricative consonants because their spectra lie mostly above 4 kHz (Fig. 5b). When fricative consonants are lost or attenuated, speech intelligibility is degraded, particularly when heard in a noisy listening environment. Thus, we chose a front-end bandwidth of 8 kHz. Thus, the sound quality of our canned speech is comparable to an FM broadcast.

Word Selection

The next important consideration is the choice of words to be stored in the Speech Waveform Library. Although Thorndike and Lorge think commonly used words can be as many as 30,000 [11], we normally use only a fraction of that number. It is interesting to note that a lengthy novel such as "Moby Dick" by Herman Melville uses only 1000 words, and "Huckleberry Finn" by Mark Twain uses only 2400 words [12]. Kucera and Francis [13] also noted that in the sample of 1,014,232 words from written texts, the 100 most common words accounted for 47.7% of the corpus; the 500 most common words accounted for 61.9%, and the 1000 most common words for 68.8%.

If verbal communication is directly related to the execution of a specific task (e.g., driving a taxi, controlling air traffic, making airline reservations, or ordering inventory parts from a catalog), the number of words used is rather limited. Once, there was a Navy study to count all the words spoken by pilots during tactical operations [14]. Among one million words counted, only 1500 different words were unique.

At any rate, the words to be stored in the speech library should be selected for each mission (i.e., surveillance, reconnaissance, etc.). Furthermore, we recommend the following steps in the selection of words to make concatenated speech more natural:

- *Singular and Plural Nouns*: For each noun, we store both singular and plural forms. For example, we store the following pairs: (*ship/ships*), (*foot/feet*), (*knife/knives*), (*child/children*), (*woman/women*), etc.

- *Five Different Forms of Verbs*: For each verb, we list five derivative forms (present, past, past perfect, present continuous, and third person singular); for example: *go, went, gone, going, goes.*

- *Homographs:* Various pronunciations of homographs (e.g., such as *rébel* and *rebél*) will be included in the Speech Waveform Library.

- *Adjective and Adverb:* For each adjective, the corresponding adverb and noun are also stored; for example: *fierce, fiercely, fierceness.*

- *Definite and Indefinite Articles:* These are difficult words to store because they are short, soft, and often coarticulated with the next word. We store three versions of both the definite and indefinite articles:

  (1) Prior to a vowel.

  (2) Prior to a consonant that begins with a voiceless stop (/p/, /t/, or /k/) or voiced stop (/b/, /d/, or /g/).

  (3) Prior to other consonant words.

  The speech waveforms corresponding to these three cases are read in sentence context, and an appropriate portion of the speech waveform is excerpted and stored.

- *Coarticulated Words*: Frequently-used coarticulated words are stored to achieve more naturalness; for example, *this is, that is, look at, did you, could you, turn on, turn off, move over,* etc.

Synonym Checking

Since the vocabulary size of the speech library is limited, it is desirable to have a method for substituting a synonym [15]. For example, assume that the speech library does not have the word "contact," but it has the word "touch." Thus, the sentence *"Are you in contact with the other team?"* can be converted to *"Are you in touch with the other team?"* at the receiver. The design of a synonym checker is closely tied to the vocabulary listing of the Speech Waveform Library.

Speaker Selection

Once words are selected, we have to select a speaker to read them. The selected speaker must have a good voice. Not all voices are equally easy to comprehend. We know from experience that some individuals have clear voices that are more easily understood even in noisy listening environments such as cocktail parties, airports, etc. This is our definition of a *good voice.* A good voice generally has the following spectral characteristics:

- *Broad Spectral Coverage*: The frequency contents of the spoken voice must cover a frequency range from 100 Hz to 8,000 Hz and beyond (Fig. 5b). The presence of high frequency components in the spoken voice is essential for creating a tonal presence.

- *Low Fundamental Pitch Frequency*: According to extensive data, low-pitch voices are more intelligible than high-pitch voices because the spectral envelope of a low-pitch voice is better defined (Fig. 7).

- *Well-Defined Resonant Frequencies*: A good voice has four to five well-defined resonant frequencies (Fig. 5b) that make clear ringing speech sounds. Such a voice has carrying power similar to the voice of a singer. In fact, good singers are trained to generate extra resonant frequencies in the high-frequency region [16].

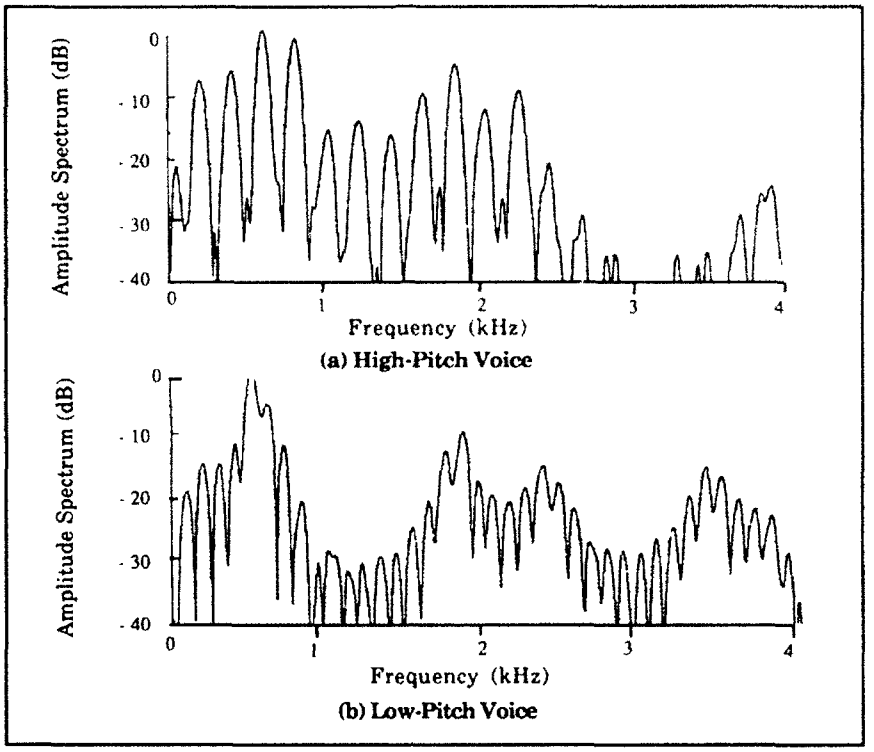


Fig. 7 — Speech spectra of low-pitch and high-pitch voices. The speech spectral envelope is better defined in the lower-pitch voice because pitch harmonics are more closely spaced.

Reading of Words

Reading of individual words is a critical element in the construction of a speech waveform library. The way we read isolated words is significantly different from the way we read the same words in a sentence context. The following rules should be followed:

- *Read fast:* Each word should be read swiftly. Otherwise, concatenated sentences will be intolerably slow. According to our investigation, a word in isolation is read approximately 20 to 40% slower than the same word in a sentence.

- *Enunciate less:* When we read an isolated word, we have more time to articulate than when we read the same word in a sentence. Thus, we should not overarticulate individual words because they will not sound natural in a sentence.

- *Maintain steady intonation:* Individual words should be read with a steady pitch to minimize pitch jitters in the concatenated sentence. We should read as many words as possible in one sitting because the intonation level of the preceding reading session will be forgotten.

- *Maintain constant loudness:* Likewise, we should maintain an equal loudness for all the words.

- *Hold microphone steady*: We recommend the use of a unidirectional dynamic microphone. It is critical to hold the microphone steady because variations in the mouth-to-microphone distance and orientation introduces fluctuation in both the speech level and low-frequency spectral content.

Preprocessing

We initially digitize speech waveforms at 48 kHz. Prior to waveform editing, we perform the following two preprocessing operations:

Three-to-One Down Sampling: Since the speech waveform was originally digitized at 48 kHz, it is subjected to a 3-to-1 down sampling to achieve a 16-kHz sampling rate. To implement a 3-to-1 downsampling operation, the speech waveform digitized at 48 kHz is low-pass filtered at 16 kHz, then the first two out of three consecutive samples are skipped. The impulse response of a sharp-cutoff low-pass filter may be obtained from the following Hamming-windowed Fourier series

$$g(i)=\begin{cases} G\left[0.54-0.46\cos\left(\dfrac{2\pi i}{I}\right)\right]\left[0.5+\displaystyle\sum_{n=1}^{N}\cos\left(\dfrac{2n\pi i}{I}-0.5\right)\right], \text{for } 0\le i\le I-1 \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

where the factor G makes the sum of the impulse response unity (i.e., a DC gain of unity). The quantity I is the total number of impulse response samples that is related to the attenuation rate beyond the cutoff frequency. The quantity N is related to the cutoff frequency for a given value of I. The impulse response is symmetric with respect to the midpoint. Thus, the phase response is linear. A 16-kHz low-pass filter with a frequency roll-off rate of approximately −180 dB per octave may be realized by letting $I = 43$ and $N = 29$ in Eq. (1). The frequency response is shown in Fig. 8.

Removal of DC Components: The DC component present in the analog-to-digital (A/D) converter output (due to equipment aging) should be removed. The presence of the DC component in the speech waveform generates audible clicks where a silent segment meets the onset speech waveform. A filter that has a zero at $z = 1$ and a pole at $z = \alpha$ ($\alpha \approx .885$) is adequate for removing the DC component present in the A/D output. The transfer function of such a filter is:

$$H(z)=\frac{1+\alpha}{2}\,\frac{1-z^{-1}}{1-\alpha z^{-1}}. \tag{2}$$

This filter has a DC gain of 0. The −3 dB cutoff frequency for $\alpha = .885$ is 156 Hz. There is no frequency ripple within the passband.

Editing of Recorded Speech

Editing isolates individual words and removes undesirable gaps before and after the word. The speech waveform is best edited by observing the waveform plot. We leave a one-millisecond gap before and after each word. The following aspects are important when editing the speech waveform:

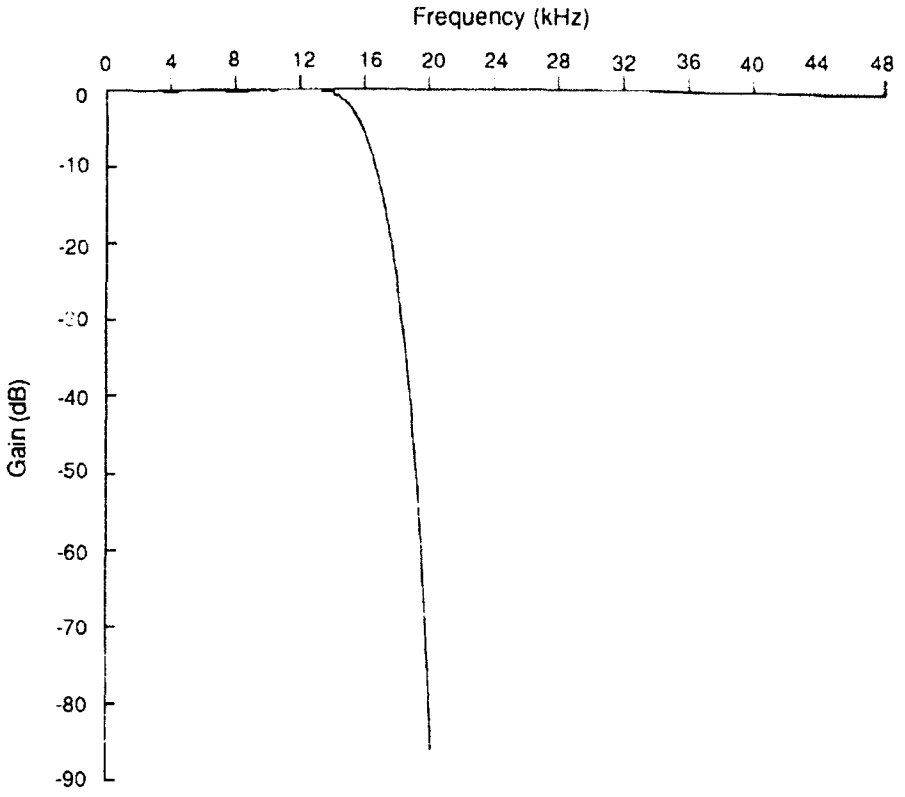


Fig. 8 — Low-pass filter frequency response for down sampling from 48 kHz to 16 kHz

- *Unvoiced Fricatives*: Unvoiced fricatives (/th/, /f/, etc.) are often not clearly visible while waveform plotting (Fig. 9). It is critical to retain those invisible speech waveforms, because speech intelligibility is highly dependent on the correct perception of a speech onset.

- *Prerelease Voicing Waveform:* On the other hand, not all visible speech waveforms are essential. When we speak a word that begins with a voiced stop consonant (/b/, /d/, /g/), we often introduce a prerelease voicing waveform (Fig. 10). It is a regular, predictable phenomenon that occurs when a voiced stop is in phrase-initial position (i.e., following silence or pause). When we speak the same word in a sentence, the prereleasing voicing waveform is considerably shorter. Thus, the prerelease voicing waveform should be shortened or completely eliminated during waveform editing.

Concatenation of Words

Earlier, at NRL, Stephanie Everett investigated a method for generating speech by concatenating segments of speech excerpted from natural speech [17]. The speech segments ranged in duration from a subphonemic size to syllable size. The total number of segments were approximately 250. Although our problem is simpler (because each word is the minimum indivisible speech unit), some of Everett's sentence-level topics in Ref. 17 (such as timing rules, intonation rules, amplitude contours, etc.) are directly applicable to our problem.

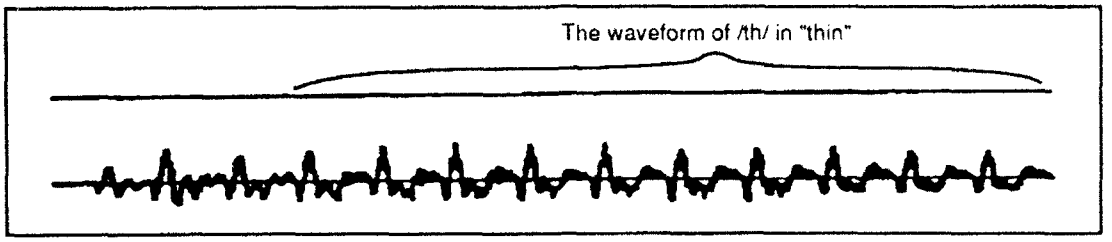


Fig. 9 — Onset of the speech waveform "thin" (approximately 100 ms). As noted, the fricative /th/ is not readily seen in this plot. Therefore, it is necessary to compute the root-mean-square value of the speech waveform to detect the onset of /th/.

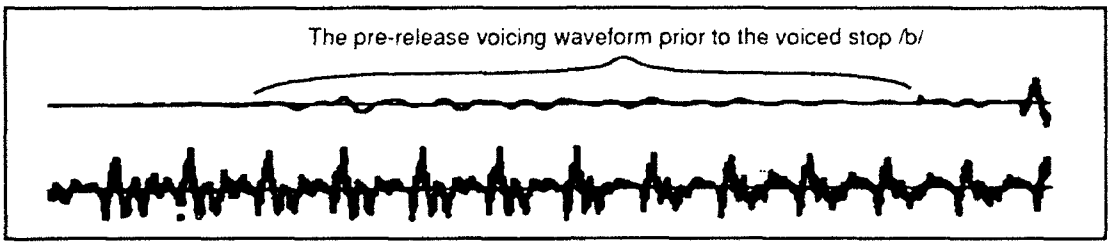


Fig. 10 — Onset of the speech waveform of "bone" (approximately 80 ms). The prerelease voicing waveform should be eliminated.

Control Variables

While concatenating the speech waveforms of individual words, we must incorporate the following three features: (1) amplitude gain for the word, (2) time gap to the next word, and (3) pitch contour within the word.

- *Amplitude gain:* We provide a set of four gains (-3dB, 0 dB, 3 dB, and 6 dB) to amplify or attenuate the speech waveform of each word. If the gain is unspecified, it is defaulted to 0 dB (i.e., no amplitude change).

- *Pause:* We provide a set of four time gaps, each a multiple of one millisecond. If the time gap is unspecified, the word is concatenated without additional time gap.

- *Pitch:* We have developed methods of varying the pitch up to 20% over the duration of a word. Changing the pitch of raw speech is not a trivial problem. For small pitch variations, we can use one of the following approaches:

  (1) *Speech Waveform Modification*: Stretching or compressing of the speech waveform can be accomplished by interpolation. This method introduces changes in both speech rate and resonant frequencies. For a small pitch change at the end of sentence, however, this method is acceptable.

  (2) *LPC Residual Waveform Modification*: The residual-excited LPC in which the prediction residual is stretched or compressed. In this method, resonant frequencies will not be affected because LPC coefficients (which control resonant frequencies) are unaltered. But speech rate will be altered as the pitch changes.

(3) *Constant-Q Transform*: The constant-Q transform has been used for both harmonic and temporal scale changes [18].

Among these methods, the Speech Waveform Modification method is simplest, and that is the one we recommend. Currently, we are working on a new technique for decoupling the pitch and utterance rates without resorting the vocoding method.

Prosody

Each language has its own distinct melodies and rhythms based on its intonation and stress patterns. People can recognize his (or her) familiar language from its melodies and rhythms even when the words are not clearly heard. Prosody is taught to persons who do not speak the native language through repeated drills and exercises [19,20]. Because the computer does not have reasoning power, we must provide the computer with a set of rigid rules to follow. But the rule is very dependent on the intention of the message. For example, "I told you to attack." This can be stated in four different ways dependent on the emphasis (Fig. 11).

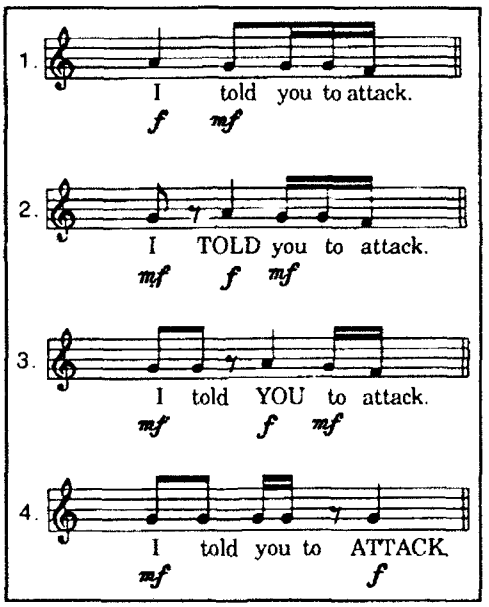


Fig. 11 — Four Possible Prosodies for "I told you to attack." The rhythm, stress, and pitch are dependent on the emphasis on the speaker's mind. (After James Martin's notation [21]).

Thus, to incorporate correct prosodies, the device or the computer must not only understand the message, but it must also know the intention of the person who originated the message. We cannot design such a smart canned speech system now or any time in the near future! Incorporation of prosodies is also required by speech synthesis by rule (see Section 1D of Ref. 5) as well as speech synthesis by concatenating speech segments (i.e., Everett's method in Ref. 17). Our problem, however, is simpler because the individual words are naturally spoken, and we are generating brief tactical messages in a manner that tactical communicators normally utter. Thus, our prosody rules are simpler.

(1) Sentence Rhythm: Time durations of individual words and pauses control the sentence rhythm. According to Lieberman who made an extensive study on these topics [22], people introduce pauses in speech for two purposes: to take a breath and to make the meaning of the words clearer. Pauses for breath are normally made at points where pauses are necessary or allowable from the point of meaning. A typical example is shown in Fig. 12. It can be seen that an appropriate location for a pause is after a verb and/or the noun after the verb.

This rule seems to be proper as evidenced in the following sentences taken from Table 1.

I need ^ air support. (^ denotes a pause)
The wind is ^ 33 knots.
Abort ^ the mission.

The length of a pause is proportional to the length of the preceding word. As stated earlier, allowable gap sizes are 0, 1, 2, and 3 ms. It is a well-known fact that the weak stressed words or syllables are spoken swiftly. Unfortunately, this feature is difficult to incorporate in canned speech, because the speech waveform cannot be shortened easily.

(2) Stress: Stress over a word is accomplished by a combination of three factors: increased loudness, raised pitch, and increased duration of the word. In the canned speech approach, the waveform cannot be stretched, but the waveform can be amplified, and pitch changes can be adjusted. If speech is a form of a command that begins with a verb (Table 1), the word to be stressed is the initial verb, because it is a critical part of the command. Otherwise, the word that precedes the pause should be stressed. For these short sentences, there will be only one stressed word. It will receive a gain of 3 to 6 dB with a pitch increase of a few percent.

(3) Intonation: Lieberman also studied intonation extensively [22]. This is a complex issue for casual speech, but it is not a major issue for brief tactical messages like those listed in Table 1. We will change intonation only for the end of the sentence. We either lower the pitch (declarative intonation) or raise the pitch (question intonation). The pitch is altered approximately 10% over the last word.

User Preference Test

We performed a survey to determine the listener preference to the following two speech generation techniques: (1) canned speech generated by the procedures outlined in this report, and (2) speech generated by a text-to-speech converter (DECtalk). We made an A/B comparison tape using ten sentences listed in Table 1. The listeners were instructed to vote for the method that they felt was best suited as a means for message transfer in tactical environments.

The twenty people participating in the test unanimously selected canned speech as being preferred or more acceptable over the speech generated by DECtalk. The listeners felt that DECtalk lacked sufficient intelligibility. One person listening to DECtalk perceived "We need ammunition" as being "We need a new mission."

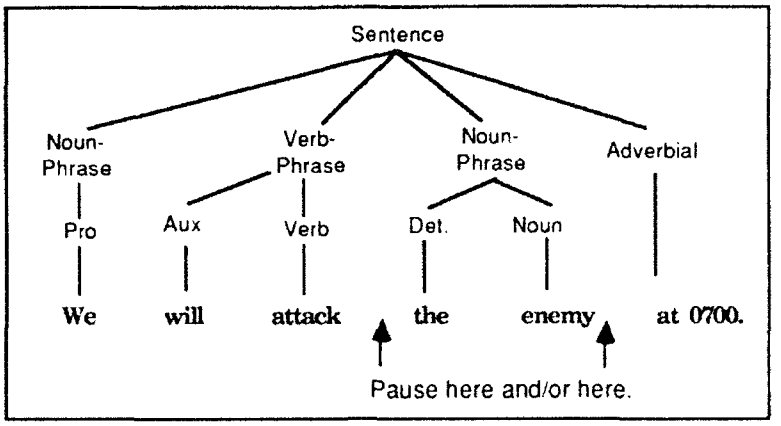


Fig. 12 — Pause Locations in a sentence (after Lieberman [22])

Table 1 — Tactical Sentences. We Generated These Sentences For Our Initial Canned Speech Experiment

Command

1. Search and destroy
2. Prepare for launch.
3. Abort the mission.
4. Observe radio silence.
5. Hold the next plane
6. Launch the tomahawk.
7. Fire for effect.
8. Make your speed one five knots.
9. Send the patrol plane off the cat.
10. Load rockets on one-two-zero.
11. Search and rescue.
12. Report to my location by 2100 hour.
13. Make your depth one five zero feet.
14. Turn on your light.
15. Steer course zero nine zero.
16. Hold enemy at 630725.
17. Attack enemy at 532714 at 0600 hour.

Status Report

18. I need air support.
19. We have no ship support.
20. Speed is kept constant.
21. You have no power.
22. Our job is done.
23. We need ammunition.
24. Number two is flying.
25. Rough terrain below.
26. All engines ahead full.
27. Weapons are tight.
28. We are moving into whiskey eight hotel.
29. There's fire in the engine room
30. This is a drill.
31. The wind is 33 knots.
32. Trains located at 543202.
33. We will return immediately.
34. Tracks show the movement of tanks.

Questions

35. What is the size of enemy force?
36. How copy?
37. Do you need air support?
38. Is your area secured?
39. Are you ready?
40. What is your location?

FORMATTED VOICE MESSAGE SYSTEM

In formatted messages, not only is vocabulary limited, but also message format is rigidly fixed. A collection of incomplete message forms are stored at both the transmitter and receiver. The computer guides the user in the selection of the proper choices to fill in this specific information. The user's choices in selecting and filling in a message correspond to indices that are the only data sent to the receiver. Since only the essential information is transmitted, the data rate is kept extremely low (well below 100 b/s).

Advantages of Formatted Messages

There are several advantages associated with the formatted message system not found in the previously discussed free-text message system. They are:

- *No Omission of Information:* There is no omission of information because the operator must fill all the information requested in the form.
- *No Extraneous Information:* On the other hand, no unessential information (particularly sensitive information) will be transmitted because they are not included in the form.
- *Easy to Read:* Since the message form is standardized, the message content is independent of the operator's writing proficiency. Therefore, the message is easy to read.
- *High Quality Output Speech:* Speech intelligibility and quality can be made as good as those of raw speech because output speech is generated by concatenating mostly phrases and sentences. In the preceding case (unformatted voice message system), we incorporated the sentence-level prosody to make output speech flow more naturally. Such is not needed in the formatted message system.
- *Extremely Low Data Rate Transmission:* Since each phrase (in certain cases, the entire sentence) is encoded, messages can be transmitted at extremely low data rates.

Three Generic Message Formats

We list three generic message formats that could be used in the Formatted Voice Message System in tactical environments.

1. Customized Messages

The message form has a number of possible statuses or conditions (Fig. 13). The operator selects an appropriate item. The system transmits only the index of the item clicked (in this case, Index 22). The speech heard at the receiver is a continuous narrative that was originally spoken as a sentence. Thus, the intelligibility of the output speech is identical to that of raw speech.

2. Message Form With Information Entries

The computer guides the user to select the proper form and to respond with the specific information. This is a questionnaire format that is familiar to most of us. Figure 14 is an example of a formatted reconnaissance report. The bold faced letters represent the information entered by the

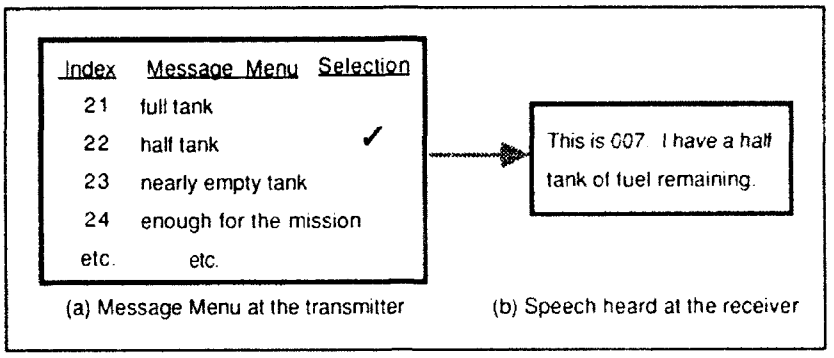


Fig. 13 — Formatted Message Type 1. The complete message is pre-made, spoken, and stored. The system has many different kinds of messages, one of which is selected from the visually-prompted message menu.

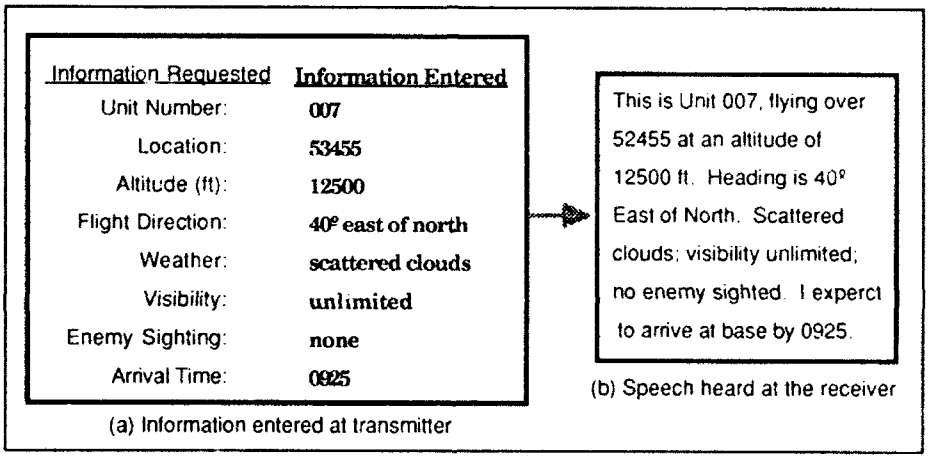


Fig. 14 — Formatted Message Type 2. This is the most familiar form where the user supplies the information requested. The transmitter transmits only the information provided.

communicator. The speech heard at the receiver is continuous narrative. The total message is approximately 20 seconds long, but the encoded information can be transmitted in approximately 1/8 second over a 2400-b/s link. The intelligibility and quality of output speech is basically as good as raw speech because all of the phrases are spoken naturally. This sort of formatted message is currently used for the telephone directory assistance by the telephone company.

3. Jargonized Messages

Tactical messages are often jargonized to make messages more easily understood in poor signal-to-noise conditions. An example is the LINK-11 voice coordination messages. LINK-11 is the digital network over which the Naval Tactical Data System (NTDS) computers communicate. The NTDS computer processes the tactical information it receives from sensors and other sources and disseminates it through LINK-11 to the other platforms in the battle group. On board ship, the NTDS Supervisor is responsible for inputting, receiving, and displaying these data, as well as for the maintenance of the link itself. The track supervisor coordinates these activities with his counterparts aboard the other

ships in the battle group. One platform's track supervisor, the Fleet Track Supervisor, is responsible for the entire network.

Because of the complexity and difficulty of maintaining LINK-11, the track supervisors communicate over the Data Systems Administration (DSA) network (Fig. 11). The DSA network is an independent, dedicated, voice radio circuit, in addition to the LINK-11 data network. In order to communicate quickly and effectively over the DSA network, track supervisors speak by using what is known as X-ray codes (Fig. 15). These are three letter brevity codes that represent all aspects of LINK-11 operations. Track supervisors can effectively coordinate link operations with a minimum of conversation using X-ray codes, a few modifier words, and some keywords. An application where the formatted voice message system could be used is to perform DSA function over the LINK-11 data channel; thus, the need for a separate DSA voice channel could be eliminated.

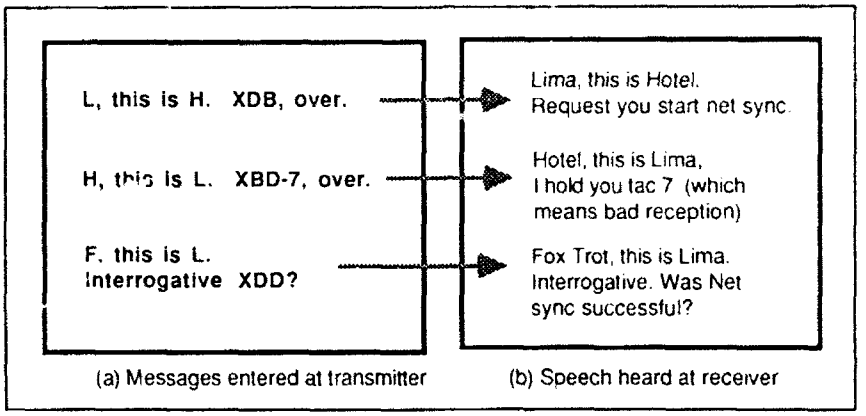


Fig. 15 — Formatted Message Type 3. These LINK-11 voice coordination messages. The messages are jargonized using X-ray code. Currently, the operator at the receiver hears the messages the way it is spoken at the transmitter. Thus, the operator must be trained to understand X-ray codes. The Formatted Message System, however, can automatically translate to plain text at the receiver. Note that the X-ray codes cited above are imitations. Real X-ray codes are classified.

There are advantages associated with eliminating the DSA voice channel. Removing the need for a suite of dedicated communication equipment would provide the Navy savings in equipment cost, as well as shipboard savings in weight, space, and power. There is also a saving of radio frequency spectrum.

Message Table

The message table contains the list of messages that can be transmitted. It must be identical at both the transmitter and receiver. At the receiver it is indexed to the speech recordings of the messages. For a formatted voice message system to be successful, the message table must be both comprehensive to transfer the intended information and concise to be manageable by the input interface.

This implies that a careful study must be undertaken to determine what messages the users need to communicate. This is a difficult task because the message table is likely to change over time. It may be a good idea to include the tools necessary for a user community to create their own messages.

These tools should include utilities to write, record, edit, and debug messages. A method is also needed for distributing messages to the communicators. The performance of the customized message system must be carefully monitored.

The physical implementation of a voice message table is not technically difficult. The main requirement is for a large amount of data storage. For example, high-quality speech has a frequency bandwidth of 8000 Hz. To digitize the speech it must be sampled at 16,000 times per second with a resolution of at least 8 bits per sample. This works out to be 128,000 bits, or 16,000 bytes for each second of stored speech. If a typical message lasts for 4 seconds and the message table contains 20 messages, approximately 1.3 Megabytes of memory are required. With the high density (over one Gigabyte) solid-state memory devices now being available, an even larger system could be implemented.

Output Interface

The output interface is the part of the system that plays out the formatted voice messages. It may also display messages as text. Each message has a corresponding speech recording. Upon receiving an index the message is looked up, and the speech is played out. The output can be directed to a loud speaker, headset, or broadcast over an intercom. If the receiving device is attached to a digital data network, the speech data could be addressed to one or more recipients and routed through the network. Since there is no processing or transmission of the speech waveform, the speech output is of the highest quality. Also, since the entire message is recorded at one time, the message sounds natural and can be understood effortlessly.

CONCLUSIONS

We examined various aspects of canned speech generation. In this approach, brief tactical messages are generated by concatenating the speech waveforms corresponding to the individual words. According to our tests, listeners unanimously preferred canned speech over synthetic speech generated by a text-to-speech converter. They selected canned speech not only for its higher intelligibility, but they also felt that canned speech was more natural. Some listeners thought that canned speech was originally recorded as sentences rather than individual words strung together.

The data rate required to transmit canned speech is very low (on the order of 20 b/s, if the vocabulary size is limited to approximately 1000 words). Such a low-data-rate voice encoding technique is essential when the network is congested. Future voice communication should be designed in such manner that no voice messages are preempted because any voice call might be delivering vital tactical information. In addition, the canned speech approach is ideally suited for transmitting speech over underwater acoustic channels.

There are other merits of canned speech. In extremely noisy environments (helicopters, tanks, etc.), live speech cannot be vocoded and transmitted because the resultant speech intelligibility is very poor. In this case, canned speech entered by a nonverbal means (e.g., keyboard) achieves a reliable message transfer. Furthermore, canned speech can readily be translated into other languages. With the availability of low-cost, high-density memory devices, canned speech generation is practical in these tactical applications where sentences are generally short and intelligibility is of primary importance.

ACKNOWLEDGMENTS

We thank Timothy McChesney and Sharon James of SPAWAR OOI for their support of this R&D effort. The authors also express thanks to Stephanie Everett and Astrid Schmidt-Nielsen of NRL who provided helpful comments.

REFERENCES

1. L. Reilly, "What Copernicus Will Mean at Sea" (Washington Technology, Feb. 1992).

2. Copernicus Project Team, "Copernicus" (Director, Space and Electronic Warfare (OP-094), ONR, Washington, DC 20350-2000).

3. G.S. Kang and L.J. Fransen, "High-Quality 800-b/s Voice Processing Algorithm," NRL Report 9301, Feb. 25, 1991.

4. G.S. Kang and L.J. Fransen, "ANDVT Rate Conversion Algorithm (from 2400 b/s to 1200 b/s)," NRL Report 9357, Dec. 27, 1991.

5. *Understanding Link-11 Guidebook for Operators, Technicians, and Net Managers*, Logicon, Inc., San Diego, CA 92121, 1990.

6. W.D. Garvey, "The Intelligibility of Speeded Speech," *J. Exp. Psychol.* **45**, 102-108 (1953).

7. D.H. Klatt, "Review of text-to-speech conversion of English," *J. Acoust. Soc. Am.* **82**(3), 737-793 (1987).

8. R. Rubinstein and H.M. Hersh, *The Human Factor* (Digital Press, Digital Equipment Corporation, 1984) p. 85.

9. T.S. Heppenheimer, "Computer Talk: Amazing new realism in synthetic speech" (Popular Science, 1984).

10. K.D. Kryter, "Validation of the Articulation Index," *J. of Acoust. Soc. of Am.* **34**(11), 1698-1702 (1962).

11. E. L. Thorndike and I. Lorge, "The Teacher's Word Book of 30,000 Words" (Teachers College, Columbia University, 1944).

12. R.J. Dixson, *Graded Exercises in English* (Simon and Schuster, New York, 1959).

13. H. Kucera and W.N. Francis, *Computational Analysis of Present-Day American English* (Brown Univ. Press, Providence, RI, 1967).

14. J.C. Webster, "Compendium of Speech Testing Material and Typical Noise Spectra for Use in Evaluating Communications Equipment" (Technical Document 191, NRAD, San Diego, CA, 1972).

15. J. Sanberg, "The Acoustics of the Singing Voice," *Scientific American*, 82-91 (1977).

16. Editors, *Webster's New Dictionary of Synonyms* (G. & S. Merriam Company, Springfield, MA, 1973).

17. S.S. Everett, "Unlimited Vocabulary Synthesis Using Line Spectrum Pairs," NRL Report 9340 (1991).

18. J.E. Youngberg, "Rate/Pitch Modification Using the Constant-Q Transform" (1979 ICASSP, 748-751, 1979).

19. English Language Service, *Stress and Intonation, Part 1 and Part 2* (The Macmillan Co., New York, New York 10022, 1967).

20. R. Ladop and C.C. Fries, *Exercise in Sound Segments, Intonation and Rhythm* (The University of Michigan Press, 1969).

21. J.G. Martin,"Rhythmic (Hierarchical) Versus Serial Structure in Speech and Other Behavior" (Psychological Review, 79(6), 487-509, 1972).

22. P. Liberman, *Intonation, Perception, and Language* (The M.I.T. Press, Cambridge, MA., 1975).